UNCLASSIFIED

Defense Technical Information Center
Compilation Part Notice

ADP010386

TITLE: Uses of the Diagnostic Rhyme Test [English Version] for Predicting the Effects of Communicators' Linguistic Backgrounds on Voice Communications in English: An Exploratory Study

DISTRIBUTION: Approved for public release, distribution unlimited

This paper is part of the following report:

TITLE: Multi-Lingual Interoperability in Speech Technology [l'Interoperabilite multilinguistique dans la technologie de la parole]

To order the complete compilation report, use: ADA387529

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, ect. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:

ADP010378 thru ADP010397

UNCLASSIFIED

USES OF THE DIAGNOSTIC RHYME TEST (ENGLISH VERSION) FOR PREDICTINGTHE EFFECTS OF COMMUNICATORS' LINGUISTIC BACKGROUNDS ON VOICE COMMUNICATIONS IN ENGLISH: AN EXPLORATORY STUDY

William D. Voiers
Dynastat, Inc.
2704 Rio Grande, Austin, TX 78705 USA
bvoiers@aol.com

ABSTRACT

Recordings of Diagnostic Rhyme Test (DRT) materials by native talkers of English (American), German and French were presented under undegraded and degraded conditions to English speaking listening crews of three national origins: American, German and French. The results were analyzed for the effects of the talker's native language, the listener's native language and all permutations of the two on scores yielded by the DRT. With undegraded speech, the total number of errors was lowest when the talkers were American, regardless of the nationality of the listeners, and when the listeners were American, regardless of the nationality of the talkers. On average, French talkers yielded the lowest DRT scores, but the interaction of talker nationality and listener nationality was significant. Errors of discrimination with respect to *voicing, sustention, sibilation* and *graveness* occurred most often.

Keywords: Intelligibility, Diagnostic Rhyme Test, multi-lingual interoperability

1. INTRODUCTION

Many factors potentially contribute to errors in speech communication in circumstances where the communi-cators are required to communicate in other than their native languages, as is frequently the case in civilian and military aviation communications. These factors include language differences in syntactical and grammatical rules. They also include differences in the phonemic alphabets of the various languages involved. Comparisons of the phonemic alphabets of the languages involved may permit identification of some of the more important sources of mis-communication, i.e., speech elements not common to the native languages of the communicators involved. Such comparisons do not, however, permit quantitative predictions regarding communication failures, nor do they permit distinctions between communication failures due to errors of articulation and those due to errors of perception — distinctions between failures due to the talker and those due to the listener.

2. PURPOSES

The purposes of this study were (1) to demonstrate the sensitivity of the Diagnostic Rhyme Test [1, 2] to the effects of communicator differences in linguistic background on voice communications conducted in English, (2) to evaluate the relative contributions of the talker's and the listener's linguistic backgrounds to voice communication failures and (3) to identify the speech elements and/or features most susceptible to misarticulation or misperception by non-native talkers of English.

3. METHODS AND MATERIALS

3.1 Speech materials

The speech materials used for this study were recordings of the test words of the Diagnostic Rhyme Test (DRT-IV). Although originally designed to aid communication scientists and engineers in pinpointing specific system defects or malfunctions, the DRT has been widely used for predicting overall intelligibility in voice communication systems and devices. It is the NATO standard and an ANSI standard for evaluating intelligibility of voice coding and communication systems and algorithms.

The DRT tests the discriminability of six distinctive features of consonant phonemes, only. It uses a 2AFC paradigm in which the listener's task with each test token or stimulus word, is to choose between two rhyming words whose initial consonants differ only with respect to one of six features: *voicing, nasality, sustention, sibilation, graveness* and *compactness*. In addition to a total score, the DRT yields more than 24 independent scores. Among these are scores for the discriminability, generally, of each feature, separate scores for each feature state, and various other subscores for each feature, e.g., separate subscores for the discriminability of *sibilation* in voiced and unvoiced phonemes.

3.2 Talkers

The talker sample consisted of three adult males from each of three linguistic backgrounds: American, German and French. They were originally recruited in their native countries by Caldwell P. Smith of the USAF Rome Air Development Center laboratory at Hanscomb AFB, Massachusetts, USA. All, presumably, had formal education in English, but their facility and experience with this language were not independently determined. Each talker recorded several randomizations of the American Diagnostic Rhyme Test words and assorted other speech materials.

3.3 Listeners

Three crews of seven test-naïve listeners, male and female, representing, respectively, American, German and French linguistic backgrounds, were also recruited from present residents of Austin, Texas. None had previous experience with the DRT. All were residing in academic or vocational

environments where English was the dominant language of everyday speech communication.

3.4 Testing procedures

The listeners were instructed in DRT testing procedures, given three practice sessions with the test and then presented recorded DRT materials by American, German and French talkers under two conditions, undegraded speech and speech masked by speech-modulated noise at an S/N of 0 dB. The speech materials were presented binaurally over TDH-39 headphones at a comfortable listening level, *circa* 79 dB SPL.

4. RESULTS

DRT results are conventionally expressed in terms of "percent correct, adjusted for chance." In a 2AFC case, the adjustment involves simply doubling the number of observed errors. We will find it convenient to adopt a system of abbreviations for denoting the various permutations of talkers' (TN) and listeners' (LN) linguistic backgrounds: A = English (American), G = German and F = French such that, e.g., GA = German talker(s) * American listener(s), FG = French talker(s)* German listener (s).

Due to the small number of talkers and listeners available for this study, the effects of "talker nationality," "listener nationality" and their interaction are statistically significant in a relatively small number of cases. However, a number of potentially important trends are strongly suggested by these results.

4.1 Results for undegraded speech

Total DRT errors for each of the nine permutations of (TN) and (LN) are shown for the undegraded case in Table 1. Scores were highest when both talkers and listeners were native-born Americans; lowest when the talkers were German and the listeners were French. Listeners of all linguistic backgrounds yielded the highest scores when the talkers were Americans, next highest when the talkers were German and lowest when the talkers were French.

Table 1. Effects of communicators' nationalities on total DRT scores

	Talkers			
Listeners	American	German	French	Mean
American	96.5	92.1	89.9	92.8
German	92.4	89.9	87.3	89.9
French	86.2	82.8	85.3	84.8
Mean	91.7	88.3	87.5	89.2

(For TN, P< .10; for LN, P< .001; for TN*LN, P< .001)

The distribution of voicing discrimination scores for the nine TN * LN permutations are shown in Table 2. Voicing scores were highest when listeners and talkers were American-born; lowest on average when the talkers were of French national origin. Overall, fewest errors occurred with German talkers; most errors occurred with French talkers.

Table 2. Effects of communicators' nationalities on discrimination scores with respect to voicing

	Talkers			
Listeners	American	German	French	Mean
American	97.0	94.9	85.1	92.4
German	90.8	94.3	78.9	88.0
French	89.6	91.7	84.5	88.6
Mean	92.7	93.6	82.8	89.6

(For TN*LN, P< .05)

As shown in Table 3, a consistent positive bias (measured as the difference between "percent correct for the positive feature state" and "percent correct for the negative feature state") appears in all cases involving French talkers, suggesting that French talkers tend to "overvoice". All listeners had a small, but statistically insignificant, tendency to perceive unvoiced phonemes as voiced when the talker was French.

Table 3. Effects of communicators' nationalities ondiscrimination biases for voicing

	Talkers			
Listeners	American	German	French	Mean
American	0.0	0.6	9.5	3.4
German	-0.6	3.0	12.5	5.0
French	-1.8	-4.8	4.8	-0.6
Mean	- 0.8	-0.4	8.9	2.6

Historically, *nasality* has proven to be the most robustly encoded of the six features dealt with by the DRT. Errors were negligible for all talker-listener permutations, but, as shown in Table 4, occurred most frequently with French listeners.

Table 4. Effects of communicators' nationalities on discrimination scores with respect to nasality

	Talkers			
Listeners	American	German	French	Mean
American	99.1	99.1	99.4	99.2
German	97.9	99.4	98.8	98.7
French	98.5	96.4	96.1	97.0
Mean	98.5	98.3	98.1	98.3

(For LN, P<.05; for LN*TN, P<.10.)

In all cases, biases with respect to nasality were less than 2%, and no distinguishing trends evident.

Results for the case of *sustention* are shown in Table 5. The main effect for LN is highly significant; the interaction LN*TN is moderately significant. No bias effects approached significance. Here as elsewhere, a significant main effect should be examined critically where an interaction involving that effect is significant. Most of the variation observed here is attributable to cases involving French listeners, the implication of which is that French listeners have greater difficulty than those of other linguistic backgrounds in

distinguishing stopped or interrupted consonants from their sustained counterparts. This phenomenon was evident independently of whether the contrasting phonemes involved were voiced (e.g. bat vs. vat) or unvoiced (e.g., pat vs. fat). However, no biases with respect to this feature approached significance.

Table 5. Effects of communicators' nationalities on discrimination scores with respect to sustention

	Talkers			
Listeners	American	German	French	Mean
American	97.6	90.5	94.0	94.0
German	90.2	83.9	89.6	87.9
French	72.0	69.3	78.3	73.2
Mean	86.6	81.2	12.8	14.9

(For LN, P<.001; for LN*TN, P<.10.)

Table 6 shows the distribution of errors with respect to *sibilation*. Errors with respect to this feature were negligible when both talkers and listeners were American, moderate for the case of American talkers and German listeners, but very frequent for all other LN * TN permutations. Moreover, the variation over

Table 6. Effects of communicators' nationalities on discrimination scores with respect to sibilation

	Talkers			
Listeners	American	German	French	Mean
American	98.5	80.4	76.2	85.0
German	93.7	78.0	72.0	81.2
French	81.8	68.8	75.9	75.5
Mean	91.3	75.7	74.7	81.6

(For LN, P<.01; for TN, P<.001; for LN*TN, P<.001.)

the nine LN * TN permutations was pronounced, both when the response options involved voiced consonants (e.g., zee vs. thee) or unvoiced consonants (e.g., sing vs. thing). For the voiced case, P<.01 for LN, P<.05 for TN and P<. 05 for the interaction, LN * TN. For the unvoiced case, P<.05 for LN, P<.05 for TN and P<.001 for LN * TN.

Sibilation bias was pronounced in the case of several LN * TN permutations. The extreme negative biases in some cases involving non-American talkers raises the possibility of recording artifacts, but the relatively small biases that occurred in the case of French listeners argues against such an explanation. Bias values for the case of *sibilation* are shown in Table 7.

Table 7. Effects of communicators' nationalities on Bias scores for sibilation

	Talkers			
Listeners	American	German	French	Mean
American	-1.8	- 25.0	-23.8	-16.9
German	-1.8	-21.4	-15.5	-12.9
French	0.6	-7.7	0.6	- 2.2
Mean	-1.0	-18.1	-12.9	-10.7

(For LN, P< .10; for LN*TN, P<.10)

Results for the "place feature," *graveness* is shown in Table 8. Although *graveness* is generally one of the most vulnerable features there is relatively little variability across LN*TN permutations except for that contributed by French listeners, who appear generally to have greatest difficulty in discriminating this feature. This difficulty is evident regardless of whether the critical consonants of the test words were voiced or unvoiced, sustained or interrupted.

Table 8. Effects of communicators' nationalities on discrimination scores with respect to graveness

	Talkers			
Listeners	American	German	French	Mean
American	87.8	90.2	88.7	88.9
German	83.9	85.1	88.1	85.7
French	70.1	77.4	80.7	79.4
Mean	83.9	84.2	85.8	84.7

((For LN, P<.001.)

Table 9 shows the distribution of biases over the nine permutations of LN and TN. Whether due to the characteristics of the talker's or to their own, listeners' responses to the grave test words were biased toward the acute state of the feature in all but two cases, both involving German talkers. This is attributable in part to the fact that four of the items on the grave subtest of the DRT require the listener to distinguish between *f* and θ, the latter of which is absent from the German phonemic alphabet.

Table 9. Effects of communicators' nationalities on biases with respect to gravèness

	Talkers			
Listeners	American	German	French	Mean
American	-16.1	6.5	-8.3	-6.0
German	-13.1	3.6	-3.6	-4.4
French	-6.5	-17.9	-12.5	-12.3
Mean	-11.9	-2.6	-8.1	- 7.6

(For LN, P < .001)

Table 10 shows the distribution of errors with respect to the place feature, *compactness*. Few errors occurred under with any permutation of LN and TN, only the LN and LN*TN effects approached statistical significance. All biases were negligible in this case.

Table 10. Effects of communicators' nationalities on total scores with respect to the feature compactness

	Talkers			
Listeners	American	German	French	Mean
American	98.8	97.3	95.8	97.3
German	97.9	98.8	96.1	97.6
French	95.2	93.5	96.4	95.0
Mean	97.3	96.5	96.1	95.5

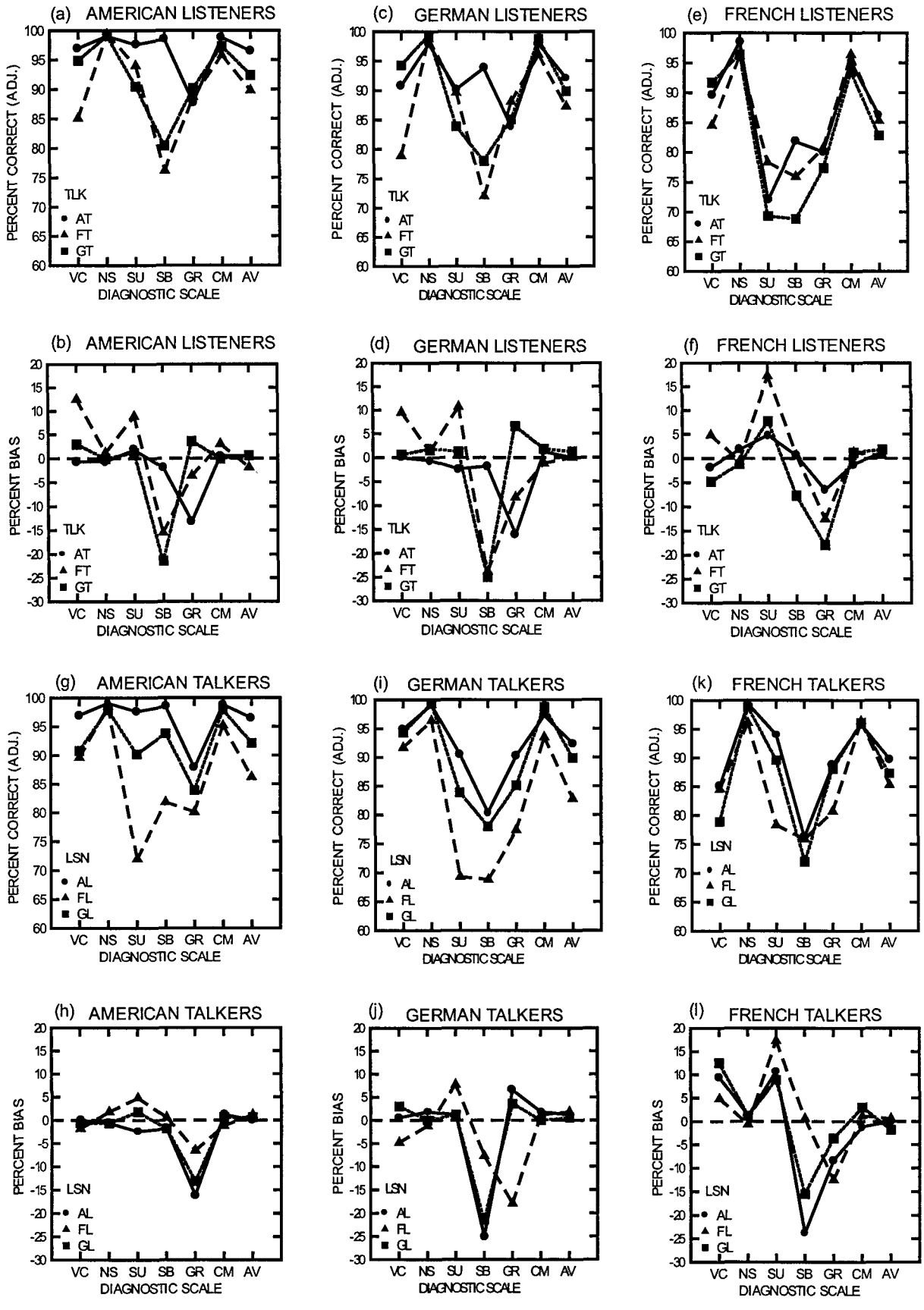

Figure 1. Diagnostic score and bias patterns for the various permutations of listener and talker nationality

4.2. Effects of speech degradation

Recordings of the DRT by the three talker samples were also presented to the three listening crews after being degraded by speech-modulated noise at a speech-to-noise ratio of 0dB. As expected, errors increased significantly across the board. The effects of degradation on total DRT errors are shown in Table 11.

Although significant in two instances, the effects of the communicators' nationalities were generally less pronounced in this case than in the case of undegraded speech, and this trend was generally maintained at the level of individual features. However, when the distribution of errors for the case of degraded speech is compared with that for undegraded speech, differences between the various LN*TN's largely disappear, as shown in Table 12. Evidently, degradation did little to potentiate communication difficulties attributable to specific LN*TN permutations.

Table 11. Effects of communicators' nationalities on total DRT scores under degraded channel conditions (0dB MNRU)

	Talkers			
Listeners	American	German	French	Mean
American	72.3	36.5	37.6	33.9
German	65.0	43.1	41.6	39.9
French	57.9	47.7	45.6	45.1
Mean	65.1	42.4	41.6	39.6

(For LN, $P < .01$; for TN, $P < .05$)

Table 12. Increase in error percentages due to speech-signal degradation

	Talkers			
Listeners	American	German	French	Mean
American	24.2	28.6	27.0	26.6
German	27.4	33.1	28.9	29.8
French	28.3	30.5	30.9	29.9
Mean	26.6	30.7	28.9	8

4.3 Relative contributions of listener nationality and talker nationality to communication failures

Figure 1 shows the results of this study from a different point of view. It permits comparisons among patterns of diagnostic scores and biases for the various LN * TN combinations.

Figure 1a shows that, for American listeners, the state of the feature, *voicing*, is most difficult to discriminate in French talkers. In both German and French talkers, *sustention* and *sibilation* are poorly discriminated. Figure 1b suggests that these difficulties are attributable to a tendency of the French talkers to "over voice" and to a tendency of both German and French talkers to "under sibilate."

Figure 1c shows that German listeners had difficulty in discriminating voicing in the case of French talkers and, otherwise, experienced difficulty in discriminating *sustention* and *sibilation* in the speech of their compatriots and that of French talkers. They exhibited a pattern of biases (Fig. 1d) similar to that of American listeners. French listeners had difficulty discriminating the states of all features except *nasality* and *compactness* in talkers of all three nationalities, including their own. They tended to perceive interrupted consonants as their sustained counterparts and to perceive grave phonemes (Fig. 1f) as their acute counterparts.

When the talkers were American, listeners of French origin had serious difficulty discriminating the states of the features *sustention*, and *sibilation*.

When talkers were of German origin, listeners of all nationalities had some difficulty discriminating *sustention*, *sibilation* and *graveness*, but French listeners had the greatest difficulty in this respect. American and German listeners exhibited pronounced negative biases with respect to *sibilation* but negligible biases in the cases of all other features. French listeners, alone, exhibited a substantial negative bias in the case of the feature, *graveness*.

When the talkers were French, listeners of all nationalities, including French, had substantial difficulty discriminating the states of *voicing* and *sibilation*. Also, French listeners had difficulty discriminating *sustention* and *graveness*. French talkers induced positive biases in *voicing* and *sustention* for listeners of all nationalities; negative biases in the cases of the feature, *sibilation*, for American and German listeners but not for their compatriots.

In the results of ANOVA described above the effects of listener nationality generally proved to be more significant than those of talker nationality. However, an examination of the data from a different point of view provides some potentially important insights. This involved comparing the nine permutations of LS *TN in terms of their error patterns over 224 items of the DRT (including 32 "easy" items. Cluster analysis was

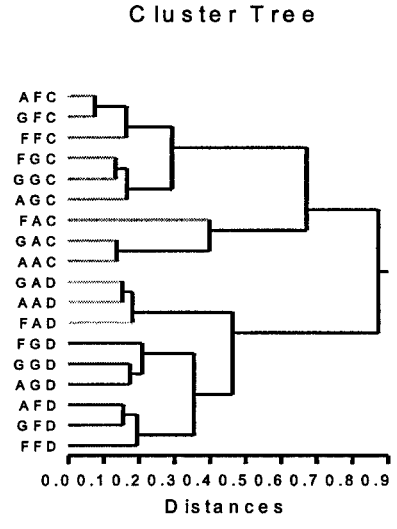

Figure 2. Cluster tree-showing similarities among LN * TN permutations with respect to error patterns across individual DRT items

the instrument of choice for this purpose. For this case, distance = Pearson r; linkage = complete.

Figure 2 shows the similarity among the nine permutations of LN and TN in terms of their error patterns under two conditions of signal quality. In the figure, the first letter of the

identifying label denotes the nationality of the listeners; the second denotes the nationality of the talkers and the third denotes the quality of the speech signal (C = clear or undegraded; D = degraded).

In the figure, there are two large clusters based on speech signal quality, one containing only the cases of undegraded speech and the other containing only cases of degraded speech. Within each of these, there are three subclusters, all of which are based on the nationality of the talkers. Thus, whereas the nationality of the listener appears to account for the bulk of communication failures, the *patterns* of these failures -- the specific types of error—appear to depend primarily on the linguistic background of the talker.

5. CONCLUSIONS

Subject to the results of additional research, the present findings suggest that remedial programs for non-native speakers of English should place primary emphasis on articulatory rather than perceptual factors in multilingual voice communications. The DRT has potential for purposes of diagnosing communication failures in circumstances requiring communication in English by non-native speakers of English. It may also be a useful tool for evaluating the efficacy of remedial training programs and for evaluating the progress of participants in such programs.

6. ACKNOWLEDGMENTS

The author is indebted to Caldwell P. Smith, whose unpublished work provided the inspiration for this study and who made available recordings of the DRT materials by French and German speakers.

7. REFERENCES

[1] Voiers, W. D. (1977) Diagnostic Evaluation of Speech Intelligibility. In M. E. Hawley (ed.) *Speech Intelligibility and Speaker Recognition.* Benchmark Papers in Acoustics, Dowden, Hutchinson and Ross, Stroudsburg, PA, USA

[2] Voiers, W. D. (1983) Evaluating Processed Speech using the Diagnostic Rhyme Test. *Speech Technology*: 30-39.